SIGNAL AND IMAGE PROCESSING FOR REMOTE SENSING

EDITED BY
C. H. CHEN

Taylor & Francis
Taylor & Francis Group
Boca Raton London New York

CRC is an imprint of the Taylor & Francis Group,
an informa business

Published in 2007 by
CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Printed in the United States of America on acid-free paper
10 9 8 7 6 5 4 3 2 1

International Standard Book Number-10: 0-8493-5091-3 (Hardcover)
International Standard Book Number-13: 978-0-8493-5091-7 (Hardcover)
Library of Congress Card Number 2006009330

Library of Congress Cataloging-in-Publication Data

Signal and image processing for remote sensing / edited by C.H. Chen.
p. cm.
Includes bibliographical references and index.
ISBN 0-8493-5091-3 (978-0-8493-5091-7)
1. Remote sensing--Data processing. 2. Image processing. 3. Signal processing. I. Chen, C. H. (Chi-hau), 1937-

G70.4.S535 2006
621.36'78--dc22 2006009330

informa
Taylor & Francis Group
is the Academic Division of Informa plc.

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

Contents

Part I Signal Processing for Remote Sensing

26

Use of the Bradley–Terry Model to Assess Uncertainty in an Error Matrix from a Hierarchical Segmentation of an ASTER Image

Alfred Stein, Gerrit Gort, and Arko Lucieer

CONTENTS

26.1 Introduction

Remotely sensed images are increasingly being used for collection of spatial information. A wide development in sensor systems has occurred during the last decades, resulting in improved spatial, temporal, and spectral resolution. The collection of data by remote sensing is generally more efficient and cheaper than by direct observation and measurement on the ground, although still of a varying quality. Data collected by sensors may be affected by atmospheric factors between sensors and the values reflected on the earth's surface, local impurities on the earth's surface, technical deficiencies of sensors and other factors. In addition, only the reflection of the sensor's signal or of the sunlight on the earth's surface is being measured, and no direct measurements are made. Consequently, the quality of maps produced by remote sensing needs to be assessed [1].

Another major issue of interest concerns the ontological interpretation of remote-sensing imagery. The truth as it supposedly exists at the earth surface is due to what one wishes to see. This can be formalized in terms of formal ontologies, although these may be subject to changes in time as well, and also a subjective interpretation is often being given.

Image classification and segmentation are important concepts in this respect. A segmentation of an image yields segments of more or less homogeneity supposed to be applicable on the earth's surface. In fact, a segmentation procedure can be done mathematically, without any more knowledge of the earth than is present in the data, hence governed by the conditions of the sensor on the earth's surface and in the atmosphere. Image classification leads to segmentation with meaningful contents. Meaningful is either what is described by the ontologies, or what a researcher aims to find. There still is a large discrepancy between the image representation on one hand and the earth's surface processes on the other hand.

After a classification is being carried out, its accuracy can be determined if ground truth is available. Classification accuracy refers to the extent to which the classified image or map corresponds with the description of a class on the earth's surface. This is commonly described by an error matrix, in which the overall accuracy and the accuracy of the individual classes are calculated. The κ-statistic is then used for testing on homogeneity [2]. Its use is somewhat controversial, because its value depends strongly on the marginal distributions [3]. Indeed, it measures the strength of agreement without taking into account the strength of disagreement. Another criticism, which we will not address in this chapter, is that it requires the presence of well-identifiable objects. In many ontologies, though, we realize that objects at the earth's surface are fuzzy or vague. In such cases, it appears to be more convenient to compare membership values.

This chapter aims to focus on an alternative to the κ-statistic. To do so, we consider on pairwise comparisons, dealing with the structure of disagreement between categories. A way to do so is by using the Bradley–Terry (BT) model [4]. Based on the logistic regression model, it may provide more details on the strength and direction of disagreement. The chapter builds on a previous study published recently [5]. It extends that paper by addressing another case study, providing more details at various places and by using a very large number of test points. We also provide additional discussion on the qualities of the BT model.

The aim of research described in this chapter is to study the use of the BT model as an alternative measure for association in the remotely sensed image for the κ-statistic. To do so, we implement and interpret a test of significance of the BT model for paired preferences in a multi-variate texture-based segmentation of geological classes from an ASTER image in Mongolia. An error matrix is obtained by validation with an existing geological map.

26.2 Concepts and Methods

A common aim in classifying a multi-spectral image is to automatically categorize all pixels in the image into classes [6]. Each pixel in the image is assigned in a Boolean fashion into one class. As a result, a thematic layer from the multi-spectral image emerges. Several classification algorithms applicable to thematic mapping from satellite images exist [7]. Supervised image classification depends on differences in the reflection on the earth's surface and hence, in bare areas, on the composition of the material on the earth's surface. Pixel-by-pixel spectral information is used as the basis for automated land-cover classification. A classification algorithm gives a classified thematic layer. Such a form of classification yields, therefore, both a segmentation of the image (i.e., different segments

emerge) and a classification (i.e., each segment has a meaning within some ontology). This is particularly the case if a supervised classification is carried out. A feature of this method, often recognized as a drawback, is that it yields a large number of very small segments, hence a highly fragmented thematic layer.

Next, the thematic layer is assessed in terms of accuracy by comparing image class samples to reference samples [8]. Representative samples are selected for each class from different locations of the image. The amount of training data usually represents between 1% and 5% of the pixels [9].

Accuracy assessment of classification is usually carried out by evaluating error matrices. An error matrix is the square contingency table. The columns represent the reference data and the rows represent the classified data [2,9]. On the basis of the error matrix, overall classification accuracy, producer's accuracy, and the user's accuracy are calculated. In addition, the κ-statistic for each individual category within the matrix can be calculated. In all these cases, the accuracy of any individual pixel is associated in a strictly Boolean fashion, that is, the classification is either correct or incorrect.

26.2.1 The κ-Statistic

The κ-statistic derived from the error matrix is a measure of agreement [2]. It is based on the difference between the actual agreement in the error matrix and the chance agreement. The sample outcome is the $\hat{\kappa}$ statistic, an estimate of κ. The actual agreement refers to the correctly classified pixel indicated by the major diagonal of the error matrix. κ measures the accuracy of classification in terms of whether the classification is better than random or not. To describe κ we assume a multinomial sampling model: each class label is derived from a finite set of possible labels. To calculate the maximum likelihood estimate $\hat{\kappa}$ of κ and a test statistic z, the error matrix is represented in mathematical form. Then $\hat{\kappa}$ is given by

$$\hat{\kappa} = \frac{p_0 - p_c}{1 - p_c} \tag{26.1}$$

where p_0 and p_c are the actual agreement and the chance agreement. Let n_{ij} equal the number of samples classified into category i, as belonging to category j in the reference data. The $\hat{\kappa}$ value can be calculated using the following formula:

$$\hat{\kappa} = \frac{n\sum_{i=1}^{k} n_{ii} - \sum_{i=1}^{k} n_{i+}n_{+i}}{n^2 - \sum_{i=1}^{k} n_{i+}n_{+i}} \tag{26.2}$$

where k is the number of classes, n_{ii} is the number of correctly classified pixels of category i, n_{i+} is the total number of pixels classified as category i, n_{+i} is the total number of actual pixels in category i, and n is the total number of pixels.

The approximate large sample variance of κ equals:

$$\text{var}(\hat{\kappa}) = \frac{1}{n}\left(\frac{\vartheta_1(1-\vartheta_1)}{(1-\vartheta_2)^2} + \frac{2(1-\vartheta_1)(2\vartheta_1\vartheta_2 - \vartheta_3)}{(1-\vartheta_2)^3} + \frac{2(1-\vartheta_2)^2(\vartheta_4 - 4\vartheta_2^2)}{(1-\vartheta_2)^4}\right) \tag{26.3}$$

where

$$\vartheta_1 = \frac{1}{n}\sum_{i=1}^{k} n_{ii}$$

$$\vartheta_2 = \frac{1}{n^2}\sum_{i=1}^{k} n_{i+}n_{+i}$$

$$\vartheta_3 = \frac{1}{n^2}\sum_{i=1}^{k} n_{ii}(n_{i+} + n_{+i})$$

$$\vartheta_4 = \frac{1}{n^3}\sum_{i=1}^{k} n_{ii}(n_{i+} + n_{+i})^2$$

Each of the terms ϑ_i can be easily calculated, and hence the κ-statistic has found its place in a wide range of (semi-)automatic classification algorithms.

Apart from being a measure of accuracy, the conditional κ coefficient κ_i is defined as the conditional agreement for the ith category. It can also be used to test the individual class (category) agreement. The maximum likelihood estimate of κ_i is given by

$$\hat{\kappa}_i = \frac{n \cdot n_{ii} - n_{i+}n_{+i}}{n \cdot n_{i+} - n_{i+}n_{+i}} \tag{26.4}$$

Its approximate large sample variance is given by

$$\text{var}(\hat{\kappa}_i) = \frac{n \cdot (n_{i+} - n_{ii})}{n_{i+}(n - n_{+i})^3}[n_{i+}n_{ii}(n_i + n_{+i} - n \cdot n_{ii}) + n \cdot n_{ii}(n_{ii} - n_{i+} - n_{+i} + n)] \tag{26.5}$$

A test of significance for a single error matrix is based on comparing the obtained classification with a random classification, that is, with assigning random labels to the individual classes at each of the test points. For this test the null hypothesis equals H_0: $\kappa = 0$, that is, the association does not significantly differ from that obtained from a random classification. The alternative hypothesis equals H_1: $\kappa \neq 0$, that is, the value of κ significantly differs from that obtained from a random classification. The test statistic equals

$$z = \frac{\hat{\kappa}}{\sqrt{\text{var}(\hat{\kappa})}} \tag{26.6}$$

A similar test of significance can be carried out for each individual category using the value of the conditional κ_i and its variance.

Different classifications may yield different error matrices. Possibly, one classification significantly improves on the other one. Comparing two error matrices, we might therefore be willing to test the null hypothesis H_0: $\kappa_1 = \kappa_2$, against the alternative hypothesis H_1: $\kappa_1 \neq \kappa_2$, that is, no significant difference occurs in association between the two error matrices. A test statistic for the significance of these two error matrices is given by

$$z_{12} = \frac{\hat{\kappa}_1 - \hat{\kappa}_2}{\sqrt{\text{var}(\hat{\kappa}_1) + \text{var}(\hat{\kappa}_2)}} \tag{26.7}$$

Computation of the z or the z_{12} statistic then allows us to evaluate H_0, as the distribution of z and that of z_{12} is approximately a normal distribution. H_0 is rejected if $z > z_{\alpha/2}$ (or

equivalently $z_{12} > z_{\alpha/2}$), where $z_{\alpha/2}$ is the $1 - \alpha/2$ confidence level of the two-tailed standard normal distribution. A similar argument applies if a one-sided test is to be carried out, for example, to investigate a claim that some new classification procedure is better than another one.

26.2.2 The BT Model

The BT model makes a pairwise comparison among n individuals, classes, or categories. It focuses on misclassifications without considering the total number and the number of correctly classified pixels (or objects). First we follow the model in this original form. Below, we address the issue of including also the total number of classified pixels. In the BT model, classes are ordered according to magnitude on the basis of misclassifications. The ordering is estimated on the basis of pairwise comparisons. Pairwise comparisons model the preference of one individual (class, category) over another [3].

The BT model has found applications in various fields, notably in sports statistics. The original paper [4] considered basketball statistics, whereas a well worked-out example on tennis appears in Ref. [3]. Both examples include multiple confrontations of various teams against one another. The BT model can be seen as the logit model for paired preference data. A logit model is a generally applicable statistical model for the relation between the probability p that an effect occurs and an explanatory variable x by using two parameters α and β. It is modeled as $\ln(p/(1-p)) = \alpha + \beta x$, ensuring that probability values are between 0 and 1 [10]. To apply the BT model to the error matrix, we consider the pair of classes C_i and C_j and let Π_{ij} denote the probability that C_i is classified as C_j and let Π_{ji} denote that C_j is classified as C_i. Obviously, $\Pi_{ji} = 1 - \Pi_{ij}$.

The BT model has parameters β_i such that

$$\text{logit}(\textstyle\prod_{ij}) = \log(\textstyle\prod_{ij}/\prod_{ji}) = \beta_i - \beta_j \tag{26.8}$$

To interpret this equation, we note that equal probabilities emerge for C_i being classified as C_j and C_j being classified as C_i if $\Pi_{ij} = \Pi_{ji} = \frac{1}{2}$, hence if $\text{logit}(\Pi_{ij}) = 0$, therefore $\beta_i = \beta_j$. If $\Pi_{ij} > \frac{1}{2}$, i.e., if C_i is more likely to be classified as C_j than C_j to be classified as C_i, then $\beta_i > \beta_j$. A value of β_i larger than that of β_j indicates a preference of misclassification of C_i to C_j above that of C_j to C_i.

By fitting this model to the error matrix, one obtains the estimates $\hat{\beta}_i$ of the β_i, and the estimated probabilities $\hat{\Pi}_{ij}$ that C_i is classified as C_j are given by

$$\hat{\Pi}_{ij} = \frac{\exp(\hat{\beta}_i - \hat{\beta}_j)}{1 + \exp(\hat{\beta}_i - \hat{\beta}_j)} \tag{26.9}$$

Similarly, from the fitted values of the BT model, we can derive fitted values for misclassification,

$$\hat{\Pi}_{ij} = \frac{\hat{\mu}_{ij}}{\hat{\mu}_{ij} - \hat{\mu}_{ji}} \tag{26.10}$$

where $\hat{\mu}_{ij}$ is the expected count value of C_i over C_j, that is, the fitted value for the model. The fitted value of the model can be derived from the output of SAS or SPSS.

26.2.3 Formulating and Testing a Hypothesis

In practice, we may be interested to formulate a test on the parameters. In particular we are interested in whether there is equality in performance of a classifier for the different classes that can be distinguished. Therefore, let H_0 be the null hypothesis: the class parameters are equal, that is, no significant difference exists between β_i and β_j. The alternative hypothesis equals H_1: $\beta_i \neq \beta_j$. To test H_0 against H_1, we compare $\hat{\beta}_i - \hat{\beta}_j$ to its asymptotic standard error (ASE). To find the ASE, we note that $\text{var}(\hat{\beta}_i - \hat{\beta}_j) = \text{var}(\hat{\beta}_i) + \text{var}(\hat{\beta}_j) - 2\ \text{cov}(\hat{\beta}_i, \hat{\beta}_j)$. Using the estimated variance–covariance matrix, each ASE is calculated as the square root of the sum of two diagonal values minus twice an off-diagonal value. An approximate 95% confidence interval for $\beta_i - \beta_j$ is then seen to be equal to $\hat{\beta}_i - \hat{\beta}_j \pm 1.96 \times \text{ASE}$. If the value 0 occurs within the confidence interval, then the difference between the two classes is not significantly different from zero, that is, no significant difference between the class parameters occurs, and H_0 is not rejected. On the other hand, if the value 0 does not occur within the confidence interval, then the difference between the two classes is significantly different from zero, indicating a significant difference between the class parameter and H_0 is rejected in favor of H_1.

26.3 Case Study

To illustrate the BT model in an actual remote-sensing study, we return to a case described in Ref. [11]. It was shown how spectral and spatial data collected by means of remote sensing can provide information about geological aspects of the earth surface. In particular in remote, barren areas, remote-sensing imagery can provide useful information on the geological constitution. If surface observations are combined with geologic knowledge and insights, geologists are able to make valid inferences about subsurface materials. The study area is located in the Dundgovi Aimag province in Southern Mongolia (longitude: 105°50′–106°26′ E and lattitude: 46°01′–46°18′ N). The total area is 1415.58 km^2. The area is characterized by an arid, mountainous-steppe zone with elevations between 1300 and 1700 m. Five geological units are distinguished: Cretaceous basalt (K1), Permian–Triassic sandstone (PT), Proterozoic granite (yPR), Triassic–Jurassic granite (yT3-J1) (an intrusive rock outcrop), and Triassic–Jurassic andesite (aT3-J1).

For the identification of general geological units we use images from the advanced spaceborne thermal emission and reflection radiometer (ASTER) satellite, acquired on 21 May, 2002. The multi-spectral ASTER data cover the visible, near infrared, shortwave and thermal infrared portions of the electromagnetic spectrum, in 14 discrete channels. Level 1B data as used in this study are radiometrically calibrated and geometrically co-registered for all ASTER bands. The combination of ASTER shortwave infrared (SWIR) bands is highly useful for the extraction of information on rock and soil types. Figure 26.1 shows a color composite of ASTER band combination 9, 6, and 4 in the SWIR range. An important aspect of this study is the validation of geological units derived with segmentation (Figure 26.2a). Reference data in the form of a geological map were obtained by expert field observation and image interpretation (Figure 26.2b).

Textural information derived from remotely sensed imagery can be helpful in the identification of geological units. These units are commonly mapped based on field observations or interpretation of aerial photographs. Geological units often show charac-

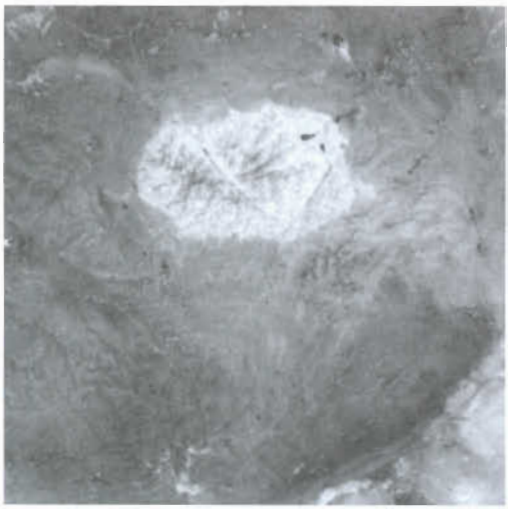

FIGURE 26.1
SWIR band combination of the ASTER image for the study area in Mongolia, showing different geological units.

teristic image texture features, for example, in the form of fracture patterns. Pixel-based classification methods might, therefore, fail to identify these units. A texture-based segmentation approach, taking into account the spatial relations between pixels, can be helpful in identifying geological units from an image scene.

For segmentation, we applied a hierarchical splitting algorithm to identify areas with homogeneous texture in the image. Similar to split-and-merge segmentation each square image block in the image is split into four sub-blocks forming a quadtree structure. The criterion used to determine if an image block is divided is based on a comparison between uncertainty of a block and uncertainty of its sub-blocks. Uncertainty is defined as the ratio between the similarity values (G-statistic), computed for an image block B, of the two most likely reference textures. This measure is also known as the confusion index (CI). The image is segmented such that uncertainty is minimized. Reference textures are defined by two-dimensional histograms of the local binary pattern and the variance texture measures. To test for similarity between an image block texture and a reference texture, the G-statistic is applied. Finally, a partition of the image with objects labeled with reference texture class labels is obtained [12].

26.3.1 The Error Matrix

Segmentation and classification resulted in a thematic layer with geological classes. Comparison of this layer with the geological map yielded the error matrix (Table 26.1). Accuracy of the overall classification equals 71.0%, and the κ-statistic equals 0.51. A major source for incorrect segmentation is caused by the differences in detail between the segmentation results and the geological map. In the map, only the main geological units are given, where the segmentation provides many more details. A majority filter of 15×15 pixels was applied to filter out the smallest objects from the ASTER segmentation map. Visually, the segmentation is similar to the geological map. However, the K1 unit is much more abundant in the segmentation map. The original image clearly shows a distinctly different texture from the surrounding area. Therefore, this area is segmented as a K1 instead of as a PT unit. The majority filtering did not provide higher accuracy values, as the total accuracy only increased by 0.5%.

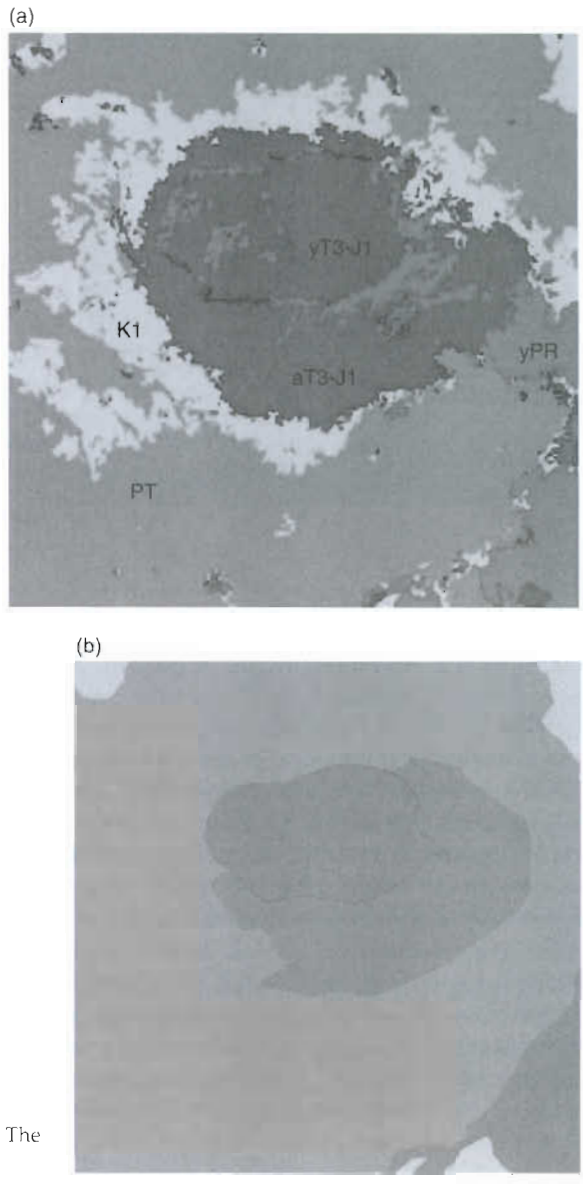

FIGURE 26.2
(a) Segmentation of the ASTER image. (b) The geological map used as a reference.

26.3.2 Implementation in SAS

The BT model was implemented as a logit model in the statistical package SAS (Table 26.2). We applied proc Genmod, using the built-in binomial probability distribution (DIST = BIN) and the logit link function. The covb option provides the estimated covariance matrix of the model parameter estimators, and estimated model parameters are obtained with the obstats option. In the data matrix, a dummy variable is set for each geological class. The variable for C_i is 1 and that for C_j is −1 if C_i is classified over C_j. The logit model has these variates as explanatory variables. Each line further lists the error value ('k') of C_i over C_j and the sum of the error values of C_i over C_j and C_j over C_i ('n'). The intercept term is excluded.

26.3.3 The BT Model

The BT model was fitted to the error matrix as shown in Table 26.1. Table 26.3 shows the parameter estimates $\hat{\beta}_i$ for each class. These values give the ranking of the category in

TABLE 26.1

Error Matrix for Classification

Classified as	Reality PT	yT3-J1	K1	aT3-J1	yPR	Total
PT	518,326	304	8,772	3,256	13,716	544,374
yT3-J1	805	86,103	0	2,836	3,848	93,592
K1	140,464	95	8,123	13,645	3,799	166,126
aT3-J1	50,825	2,997	827	95,244	2,621	152,514
yPR	23,788	8,082	60	21,226	31,569	84,725
Total	734,208	97,581	17,782	136,207	55,553	1,041,331

Note: PT (Permian and Triassic formation), yT3-J1 (upper Triassic and lower Jurassic granite), K1 (lower Cretaceous basalt), aT3-J1 (upper Triassic and lower Jurassic andesite) and yPR (Penterozoic granite). The overall classification accuracy is 71.0%, the overall κ-statistic equals 0.51.

comparison with the reference category. Table 26.3 also shows standard errors for the $\hat{\beta}_i$s for each class. The β parameter for class yPR is not estimated, being the last in the input series, and is hence set equal to 0. The highest $\hat{\beta}$ coefficient equal to 1.545 is observed for class PT (Permian and Triassic formation) and the lowest value is equal to –1.484 for class K1 (lower Cretaceous basalt). Standard errors are relatively small (below 0.016), indicating that all coefficients significantly differ from 0. Hence the significantly highest erroneous classification occurs for the class PT and the lowest for class K1.

Estimated $\hat{\beta}_i$ values are used in turn to determine misclassification of one class over another (Table 26.4). This anti-symmetric matrix shows again relatively high values for differences of geological classes with PT and lower values for differences with class K1. Next, the probability of a misclassification is calculated using Equation 26.9 (Table 26.5). For example, the estimate of probability of misclassifying PT as upper Triassic and lower Jurassic granite (yT3-J1) is 0.28, whereas that of a misclassification of yT3-J1 as PT equals

TABLE 26.2

```
data matrix;
input PT yT3J1 K1 aT3J1 yPR k1 n1 k2 n2 w;
k = (k1/n1);
n = (k1/n1 + k2/n2);
cards;
```

Data Matrix	Input	PT	yT3-J1	K1	aT3-J1	yPR *k n*; Cards
–1	1	0	0	0	304	1,109
–1	0	1	0	0	8,772	149,236
–1	0	0	1	0	3,256	54,081
–1	0	0	0	1	13,716	37,504
0	–1	1	0	0	0	95
0	–1	0	1	0	2,836	5,833
0	–1	0	0	1	3,848	11,930
0	0	–1	1	0	13,645	14,472
0	0	–1	0	1	3,799	3,859
0	0	0	–1	1	2,621	23,847

```
proc genmod data = matrix;
model k/n = PT yT3J1 K1 aT3J1 yPR/NOINT DIST = BIN link = logit covb
obstats;
run;
```

TABLE 26.3

Estimated Parameters $\hat{\beta}_i$ and Standard Deviations $\text{se}(\hat{\beta}_i)$ for the BT Model

	PT	yT3-J1	K1	aT3-J1	yPR
$\hat{\beta}_i$	1.545	0.612	−1.484	0.317	0.000
$\text{se}(\hat{\beta}_i)$	0.010	0.016	0.014	0.010	0.000

Note: See Table 26.1.

TABLE 26.4

Values for $\hat{\beta}_i - \hat{\beta}_j$ Comparing Differences of One Geological Class Over Another

	PT	yT3-J1	K1	aT3-J1	yPR
PT	0.000	−0.932	−3.029	−1.228	−1.545
yT3-J1	0.932	0.000	−2.097	−0.295	−0.612
K1	3.029	2.097	0.000	1.801	1.484
aT3-J1	1.228	0.295	−1.801	0.000	−0.317
yPR	1.545	0.612	−1.484	0.317	0.000

Note: See Table 26.1.

0.72: it is therefore much more likely that PT is misclassified as yT3-J1, than yT3-J1 as PT. Table 26.6 shows the observed and fitted entries in the error matrix (in brackets) for the BT model. We notice that the estimated and observed differences are relatively close. An exception is the expected entry for the yPR–aT3-J1 combination, where fitting is clearly erroneous.

To test the significance of differences, ASEs are calculated, yielding a symmetric matrix (Table 26.7). Low values (less than 0.02) emerge for the different class combinations, mainly because of the large number of pixels. Next, we test for the significance of differences (Table 26.8), with H_0 being equal to the hypothesis that $\hat{\beta}_i = \hat{\beta}_j$, and the alternative hypothesis H_1 that $\hat{\beta}_i \neq \beta_j$. Because of the large number of pixels, H_0 is rejected for all class combinations. This means that there is a significant difference between the parameter values for each of these classes.

We now turn to the standardized data, where also the diagonal values of the error matrix are included in the calculations.

TABLE 26.5

Estimated Probabilities $\hat{\Pi}_{ij}$ Using Equation 26.9 of Misclassifying C_i over C_j Using Standardized Data

	PT	yT3-J1	K1	aT3-J1	yPR
PT		0.28	0.05	0.23	0.18
yT3-J1	0.72		0.11	0.43	0.35
K1	0.95	0.89		0.86	0.82
aT3-J1	0.77	0.57	0.14		0.42
yPR	0.82	0.65	0.18	0.58	

Note: See Table 26.1.

TABLE 26.6

Observed and Expected Entries in the Error Matrix

	PT	yT3-J1	K1	aT3-J1	yPR
PT		304	8,772	3,256	13,716
		(313)	(6,885)	(12,255)	(6,595)
yT3-J1	805		0	2,836	3,848
	(796)		(10)	(2,489)	(4,194)
K1	140,464	95		13,645	3,799
	(142,351)	(85)		(12,421)	(3,146)
aT3-J1	50,825	2,997	827		2,621
	(41,826)	(3,344)	(2,051)		(41,240)
yPR	23,788	8,082	60	21,226	
	(30,909)	(7,736)	(713)	(56,625)	

Note: See Table 26.1.

TABLE 26.7

Asymptotic Standard Errors for the Parameters in the BT Model, Using the Standard Deviations and the Covariances between the Parameters

	PT	yT3-J1	K1	aT3-J1	yPR
PT		0.017	0.011	0.008	0.010
yT3-J1	0.017		0.019	0.016	0.016
K1	0.011	0.019		0.012	0.014
aT3-J1	0.008	0.016	0.012		0.010
yPR	0.010	0.016	0.014	0.010	

Note: See Table 26.1.

TABLE 26.8

Test of Significances in Differences between Two Classes

		$\hat{\beta}_i - \hat{\beta}_j$	ASE	*t*-Ratio	H_0
PT	yT-J1	−0.932	0.017	−54.84	Reject
	K1	−3.029	0.011	−280.99	Reject
	aT-J1	−2.097	0.019	−108.02	Reject
	yPr	−1.228	0.008	−145.39	Reject
yT3-J1	K1	−0.295	0.016	−18.08	Reject
	aT-J1	1.801	0.012	145.01	Reject
	yPr	−1.545	0.010	−155.96	Reject
K1	aT-J1	−0.612	0.016	−39.47	Reject
	yPr	1.484	0.014	108.69	Reject
aT3-J1	yPr	−0.317	0.010	−32.77	Reject

Note: See Table 26.1.

26.3.4 The BT Model for Standardized Data

When applying the BT model, diagonal values are excluded. This may have the following effect. If C_1 contains k_1 pixels identified as C_2, and C_2 contains k_2 pixels identified as C_1, then the BT model analyzes the binomial fraction $k_1/(k_1+k_2)$. The denominator $k_1 + k_2$, however, depends also on the total number of incorrectly identified pixels in the two classes, say n_1 for C_1 and n_2 for C_2. Indeed, suppose that the number of classified pixels doubles for one particular class. One would have the assumption that this does not affect the misclassification probability. This does not apply to the BT model as presented above. As a solution, we standardized the counts per row by defining $k = (k_1/n_1)/(k_1/n_1 + k_2/n_2)$ and $n = 1$. This has as an effect that the diagonals in the error matrix are scaled to 1, and that matrix elements not on the diagonal contain the number of misclassified pixels for each class combination. Furthermore, a large number of classified pixels should lead to lower standard deviations than a low number of classified pixels. To take this into account, a weighting was done, with weights equal to $k_1 + k_2$ being equal to the number of incorrect classifications.

Again, a generalized model is defined, with as a dependent variable the ratio k/n and the same explanatory variables as in Table 26.2. The SAS input file is given in Table 26.9. The first five columns, describing particular class combinations, are similar to those in Table 26.2, and the final five columns are described above. The variables in the model have also been described in Table 26.2. Estimated parameters $\hat{\beta}_i$, together with their standard deviations are given in Table 26.10. We note that the largest $\hat{\beta}_i$ value again occurs for the class PT (Permian and Triassic formation) and the lowest value for class K1 (lower Cretaceous basalt). Class PT therefore has the highest probability of being misclassified and class K1 has the lowest probability. Being negative, this probability is smaller than that for Penterozoic granite (yPR). In contrast to the first analysis, none of the parameters significantly differs from 0, as is shown by the relatively large standard deviations.

Subsequent calculations are carried out to compare differences in classes, as was done earlier. For example, misclassification probabilities corresponding to Table 26.5 are now in Table 26.11 and observed and expected entries in the error matrix are in Table 26.12. We now observe first that some modeled values are extremely good (such as the combinations between PT and yT3-J1) and second, some modeled values are totally different

TABLE 26.9

```
data matrix;
input PT yT3J1 K1 aT3J1 yPR k1 n1 k2 n2 w;
k = (k1/n1);
n = (k1/n1 + k2/n2);
cards;
```

−1	1	0	0	0	304	518,326	805	86,103	1,109
−1	0	1	0	0	8,772	518,326	140,464	8,123	149,236
−1	0	0	1	0	3,256	518,326	50,825	95,244	54,081
−1	0	0	0	1	13,716	518,326	23,788	31,569	37,504
0	−1	1	0	0	0	86,103	95	8,123	95
0	−1	0	1	0	2,836	86,103	2,997	95,244	5,833
0	−1	0	0	1	3,848	86,103	8,082	31,569	11,930
0	0	−1	1	0	13,645	8,123	827	95,244	14,472
0	0	−1	0	1	3,799	8,123	60	31,569	3,859
0	0	0	−1	1	2,621	95,244	21,226	31,569	23,847

```
proc genmod data = matrix;
model k/n = PT yT3J1 K1 aT3J1 yPR/NOINT DIST = BIN covb obstats;
run;
```

TABLE 26.10

Estimated Parameters $\hat{\beta}_i$ and Standard Deviations se($\hat{\beta}_i$) Using the Weighted BT Model for Standardized Data

	PT	yT3-J1	K1	aT3-J1	yPR
$\hat{\beta}_i$	4.813	1.887	−3.337	2.245	0.000
se($\hat{\beta}_i$)	5.404	4.585	6.297	3.483	0.000

TABLE 26.11

Estimated Probabilities $\hat{\Pi}_{ij}$ Using Equation 26.9 of Misclassifying C_i over C_j Using Standardized Data

	PT	yT3-J1	K1	aT3-J1	yPR
PT		0.05	0.00	0.07	0.01
yT3-J1	0.95		0.01	0.59	0.13
K1	1.00	0.99		1.00	0.97
aT3-J1	0.93	0.41	0.00		0.10
yPR	0.99	0.87	0.03	0.90	

from the observed values, in particular, negative values still emerge. These differences occur in particular in the right corner of the matrix. Testing of differences between different classes (Table 26.13) can again be carried out in a similar way, with again significance occurring in all differences, due to the very large number of pixels.

26.4 Discussion

This chapter focuses on the BT model for summarizing the error matrix. The model is applied to the error matrix, derived from a segmentation of an image using a hierarchical

TABLE 26.12

Observed and Expected Entries in the Error Matrix

	PT	yT3-J1	K1	aT3-J1	yPR
PT		304	8772	3256	13716
		(304)	(117179)		(22340)
yT3-J1	805		0	2836	3848
	(804)		(0)	(2780)	(4005)
K1	140464	95		13645	3799
	(10515)	(0)		(9342)	*
aT3-J1	50825	2997	827		2621
	*	(3057)	(1208)		*
yPR	23788	8082	60	21226	
	(14605)	(7766)	*	*	

Note: See Table 26.1; entries marked * indicate an estimate of a negative value.

TABLE 26.13

Test of Significances in Differences between Two Classes Using Standardized Data

		$\hat{\beta}_i - \hat{\beta}_j$	ASE	*t*-Ratio	H_0
PT	yT3-J1	−2.926	0.017	−172.09	Reject
	K1	−8.150	0.011	−756.06	Reject
	aT3-J1	−5.224	0.019	−269.15	Reject
	yPR	−2.568	0.008	−304.09	Reject
yT3-J1	K1	0.358	0.016	21.94	Reject
	aT3-J1	5.582	0.012	449.40	Reject
	yPR	−4.813	0.010	−485.93	Reject
K1	aT3-J1	−1.887	0.016	−121.64	Reject
	yPR	3.337	0.014	244.36	Reject
aT3-J1	yPR	−2.245	0.010	−232.07	Reject

segmentation algorithm. The image is classified on geological units. The κ-statistic, measuring the accuracy of the whole error matrix, considers the actual agreement and chance agreement, but ignores asymmetry in the matrix. In addition, the conditional κ-statistic measures the accuracy of agreement within each category, but does not consider the preference of one category over another category. These measures of accuracy only consider the agreement of classified pixels and reference pixels. In this study we extended these measures with those from the BT model to include chance agreement and disagreement. The BT model in its original form studies preference of one category over another. A pairwise comparison between classes gives additional parameters as compared to other measures of accuracy. The model also yields expected against observed values, estimated parameters, and probabilities of misclassification of one category over another category.

Using the BT model, we can determine both the agreement within a category and disagreement in relation to another category. Parameters computed from this model can be tested for statistical significance. This analysis does not take into account the categories with zero values in combination with other categories. A formal testing procedure can be implemented, using ASEs. The class parameters β_i provide a ranking of the categories. The BT model shows that a class, which is easier to recognize, is less confused with other classes.

At this stage it is difficult to say which of the two implemented BT models is most useful and appropriate for application. It appears that on the one hand the standardized model allows us to give an interpretation that is fairer and more stable in the long run, but the sometimes highly erroneous estimates of misclassification are a major drawback for application. The original, unstandardized BT model may be applicable for an error matrix as demonstrated in this study, but a large part of the available information is ignored. One reason is that the error matrix is typically different from the error matrix as applied in Ref. [11].

Positional and thematic accuracy of the reference data is crucial for a successful accuracy assessment. Often the positional and thematic errors in the reference data are unknown or are not taken into account. Vagueness in the class definition and the spatial extent of objects is not included in most accuracy assessments. To take into account uncertainty in accuracy assessment, membership values or error index values could be used.

26.5 Conclusions

We conclude that the BT model can be used for a statistical analysis of an error matrix obtained by a hierarchical classification of a remotely sensed image. This model relies on the key assumption that misclassification of one class as another is one minus the probability of misclassifying the other class as the first class. The model provides parameters and estimates for differences between classes. As such, it may serve as an extension to the κ-statistic. As this study has shown, more directed information is obtained, including a statement whether these differences are significantly different from zero.

References

1. Foody, M.G., 2002, Status of land cover classification accuracy assessment, *Rem. Sens. Environ.*, 80, 185–201.
2. Congalton, R.G., 1994, A Review of assessing the accuracy of classifications of remotely sensed data, in *Remote Sensing Thematic Accuracy Assessment: A compendium*, Fenstermaker, K.L., ed., *Am. Soc. Photogramm. Rem. Sens.*, Bethseda, pp. 73–96.
3. Agresti, A., 1996, *An Introduction to Categorical Data Analysis*, John Wiley & Sons, Inc., New York.
4. Bradley, R.A. and Terry, M.E., 1952, Rank analysis of incomplete block designs I. The method of paired comparisons, *Biometrika*, 39, 324–345.
5. Stein, A., Aryal, J., and Gort, G., 2005, Use of the Bradley–Terry model to quantify association in remotely sensed images, *IEEE Trans. Geosci. Rem. Sens.*, 43, 852–856.
6. Lillesand, T.M. and Kiefer, R.W., 2000, *Remote sensing and Image Interpretation*, 4th edition, John Wiley & Sons, Inc., New York.
7. Gordon, A.D., 1980, *Classification*, Chapman & Hall, London.
8. Janssen, L.L.F. and Gorte, B.G.H., 2001, Digital image classification, in *Principles of Remote Sensing*, L.L.F. Janssen and G.C. Huurneman, eds., 2nd edition, ITC, Enschede, pp. 73–96.
9. Richards, J.A. and Jia, X., 1999, *Remote Sensing Digital Image Analysis*, 3rd edition, Springer-Verlag, Berlin.
10. Hosmer, D.W. and Lemeshow, S., 1989, *Applied Logistic Regression*, John Wiley & Sons, Inc., New York.
11. Lucieer, L., Tsolmongerel, O., and Stein, A., 2004, Texture-based segmentation for identification of geological units in remotely sensed imagery, in *Proc. ISSDQ '04*, A.U. Frank and E. Grum, eds., pp. 117–120.
12. Lucieer, A., Stein, A., and Fisher, P., 2005, Texture-based segmentation of high-resolution remotely sensed imagery for identification of fuzzy objects, *Int. J. Rem. Sens.*, 26, 2917–2936.